Who's your neighbor? New computational approaches for functional genomics

Michael Y. Galperin and Eugene V. Koonin*

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20894, USA *Corresponding author (koonin@ncbi.nlm.nih.gov)*

Received 20 December 1999; accepted 18 April 2000

Several recently developed computational approaches in comparative genomics go beyond sequence comparison. By analyzing phylogenetic profiles of protein families, domain fusions, gene adjacency in genomes, and expression patterns, these methods predict many functional interactions between proteins and help deduce specific functions for numerous proteins. Although some of the resultant predictions may not be highly specific, these developments herald a new era in genomics in which the benefits of comparative analysis of the rapidly growing collection of complete genomes will become increasingly obvious.

Keywords; protein function prediction, gene sequence, Rosetta Stone, clusters of orthologs, phylogenetic patterns

The analysis of the first several bacterial, archaeal, and eukaryotic genomes to be sequenced proved to be an exciting, but also a humbling, exercise. The primary methodology applied to these genome sequences involved database search using sequence comparison programs, such as BLAST, and subsequently methods that allow the detection of subtle sequence conservation, such as PSI-BLAST or HMMer[1–3]. The good news from this type of analysis was that the majority of the proteins encoded in each of these genomes, between 70% and 90%, have homologs in distant species, giving us hope that, at least in principle, the genomes should be interpretable. At the same time, the results clearly show how little we actually know about even the simplest cells. Indeed, sequence comparison methods, even the best ones, are of little help when a protein has no homologs in current databases or when all database hits are to uncharacterized gene products from other sequenced genomes.

The individual estimates vary widely, but on average, there is no clear functional prediction for at least 30–35% of genes in most genomes, and for many of the rest, only general predictions can be made. Now that the count of completely sequenced genomes has exceeded 30 (ref. 4), and many more, including the human genome, are in the pipeline[5,6], one cannot help asking: Are there ways by which comparative genomics could help functional prediction reach beyond what can be achieved by straightforward database search? In this review, we discuss recent developments in sequence analysis of multiple complete genomes that, when analyzed simultaneously from different angles, offer qualitatively new opportunities for predicting gene functions in each of them.

Orthologous families

Conceptually, all these new approaches may be conveniently brought together through the notion of neighborhood, or context[7]. Several types of neighbor relationships can be usefully applied to genes (see Fig. 1). The one that departs the least from the traditional sequence similarity analysis is clustering in the phylogenetic space, or orthology. Orthologs are genes that are connected by vertical evolutionary descent ("the same" gene in different species) as opposed to paralogs, which are genes related by duplication within a genome[8,9].

Typically, orthologs perform the same function; therefore, delineation of orthologous families from a wide range of species justifies transfer of functional annotation. The major complication with this approach is that orthology is not necessarily a one-to-one relationship because a single gene in one phylogenetic lineage may correspond to a whole family of paralogs in another lineage[10]. For such one-to-many and many-to-many relationships, transfer of functional assignments requires more caution because some of the paralogs could have acquired new functions.

A remarkable result of the recent systematic analysis of orthologous families from all completely sequenced genomes is that 60–80% of bacterial and archaeal genes belong to about 2100 clusters of orthologous groups of proteins (COGs[10]), each of which includes orthologs from at least three phylogenetically distant species[10–12]. Thus, in principle, characterization of only 2500 or so genes (even considering the uncertainty due to paralogy) could take us a long way toward understanding the functional layout of prokaryotic genomes, or at least their conserved core.

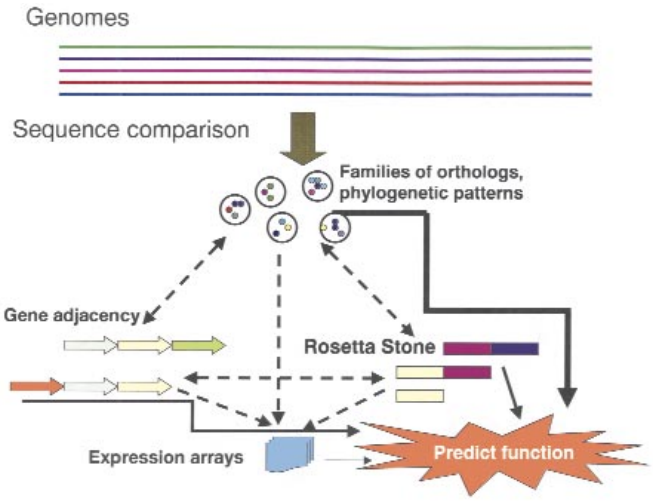

Figure 1. Context-based approaches in comparative genomics. Broken arrows indicate information flow between different types of data. Solid arrows show contribution of different types of analysis to functional prediction; very loosely, the thickness of the lines shows our evaluation of their relative significance.

Table 1. Complementary phylogenetic patterns, non-orthologous displacement and prediction of protein functions

	Species[a]																	
	Archaea				Eukaryota		Bacteria											
	Af	Mj	Mth	Ph	Sc	Aa	Tm	Ssp	Ec	Bs	Mt	Hi	Hp	Mgp	Bb	Tp	Ctp	Rp
Pathway/ Enzyme																		
Translation[b]																		
Class II lysyl-tRNA synthetase (COG1190)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	+	+	-
Class I lysyl-tRNA synthetase (COG1384)	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+
Glycolysis[c]																		
FBA (COG0191)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-
DhnA-type FBA (COG1830)	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-	+	-
Thymidylate[d] biosynthesis																		
Thymidylate synthase (COG0207)	+	+	+	-	+	-	-	-	+	+	+	+	-	+	-	-	-	-
Predicted novel thymidylate synthase (COG1531)	-	-	-	+	-	+	+	+	-	-	+	-	+	-	+	+	+	+

[a]Aa, *Aquifex aeolicus*, Af, *Archaeoglobus fulgidus*, Bb, *Borrelia burgdorferi*, Bs, *Bacillus subtilis*, Ctp, *Chlamydia trachomatis & pneumoniae*, Ec, *Escherichia coli*, Hi, *Haemophilus influenzae*, Hp, *Helicobacter pylori*, Mgp, *Mycoplasma genitalium & pneumoniae*, Mj, *Methanococcus jannaschii*, Mth, *Methanobacterium thermoautotrophicum*, Ph, *Pyrococcus horikoshii*, Rpr, *Rickettsia prowazekii*, Sce, *Saccharomyces cerevisiae*, Ssp, *Synechocystis sp.*, Tm, *Thermotoga maritima*, Tp, *Treponema pallidum*. [b]In Tp, the only functional lysyl-tRNA synthetase is probably the class I enzyme; the class II enzyme is a distinct truncated form that is likely to have a function other than translation. [c]Aa and Ec possess both types of FBA; Rp lacks glycolysis. [d]Mt is predicted to posses both types of thimydylate synthase

Examination of the annotations that are associated, in the GenBank database, with the proteins comprising the COGs shows that orthologs are often annotated differently. Among the 786 COGs with no paralogs, 194 include proteins with conflicting annotations, and 83 more consist of "hypothetical" proteins whose function could be predicted from detailed sequence analysis[12]. Thus, the conceptually straightforward step of identifying families of orthologs has the potential of significantly improving the depth and coherence of functional annotations in public databases.

The concept of orthologous families can be enhanced through the analysis of phylogenetic patterns or profiles[10,13,14]. The phylogenetic pattern for each family of orthologs is defined by the set of genomes in which the family is represented. Genes that function at different steps of the same pathway frequently have the same phylogenetic profile. Thus, the concept of phylogenetic patterns may have certain predictive power—groups of genes with the same phylogenetic profile are more likely to be functionally connected than those with different profiles. For example, only 81 COGs among 2100 are universal (i.e., represented in all completely sequenced genomes) and 56 of these consist of proteins whose functions are related to translation. So it might not be unreasonable to hypothesize that some of the universal COGs whose functions are not yet known also have a translation-related role, particularly if this is compatible with additional evidence. For example, the universal COG0012 (prototyped by the *Escherichia coli* protein YchF) consists of proteins that, by sequence analysis, are predicted to possess GTPase activity. Furthermore, clustering of the proteins in this COG by sequence similarity grouped together archaeal and eukaryotic proteins, which is typical of translation machinery components[15,16]. As many GTPases are involved in translation, it is a tempting and testable speculation that COG0012 consists of yet uncharacterized translation factors.

Non-orthologous gene displacement

The distribution of phylogenetic patterns, however, is dramatically confounded by such major evolutionary phenomena as partial redundancy in gene functions, non-orthologous gene displacement (NOGD), horizontal gene transfer, and lineage-specific gene loss (see "Comparative genomics glossary"). Indeed, because of these effects, the 2100 COGs encompass as many as 1229 distinct phylogenetic patterns. Furthermore, with the exception of the core translation machinery, several RNA polymerase subunits, and a few components of the molecular chaperone apparatus, no other systems and pathways are based upon ubiquitous proteins, and moreover, most are not characterized by a single phylogenetic pattern. Even in such central biochemical pathways as glycolysis and the TCA cycle, major variations are seen, resulting both from modification of the pathways themselves and from NOGD[1,17,18].

Occurrence of NOGD in essential functions can be explored systematically by detecting complementary, rather than identical or similar, phylogenetic patterns. The complementarity, however, is unlikely to be perfect because of partial functional redundancy—some organisms, particularly those with larger genomes, may encode both proteins providing the given function.

Table 1 shows three examples of partial complementarity between phylogenetic patterns that can be used to detect NOGD and to predict protein functions. The case of the two unrelated lysyl-tRNA synthetases has become an epitome of NOGD[19–21]. In this case, the phylogenetic patterns are nearly perfectly complementary, the

Comparative genomics glossary

Definitions of some evolutionary phenomena that are critical for comparative genomics[16,39–42]:

- *Redundancy of gene function*—most, if not all, genomes encode two or even more proteins that can perform certain functions, making each of the respective genes non-essential. There is an inverse correlation between the level of redundancy and the total number of genes.
- *Non-orthologous gene displacement (NOGD)*—displacement, in the course of evolution, of a gene coding for a protein responsible for a particular function with a non-orthologous (unrelated or distantly related) but functionally analogous gene.
- *Horizontal gene transfer*—transfer of genes from one phylogenetic lineage to another, in some cases, distant one. Comparative analysis of sequenced genomes suggests that horizontal gene transfer is common in evolution and involves a significant fraction of genes, at least in prokaryotes. Rigorous proof of horizontal transfer for many suspect genes, however, may be difficult.
- *Lineage-specific gene loss*—elimination of an ancestral gene in a particular phylogenetic lineage. The underlying mechanisms may include both actual deletion and rapid evolution that changes a gene's identity beyond recognition. In the evolution of prokaryotes, lineage-specific gene loss may be as widespread as horizontal gene transfer. Together, these phenomena are most likely to account for the wide diversity of phylogenetic patterns seen among orthologous gene families.

only exception being the presence of both types in the spirochete *Treponema pallidum* (Table 1). The unexpected discovery of the class I lysyl-tRNA synthetase in archaea and spirochetes had been made before large-scale comparative genome analysis have become possible[19,20]. In retrospect, it seems that the complementarity of the phylogenetic patterns of the typical class II lysyl-tRNA synthetase and an unassigned class I synthetase would have been sufficient for a functional prediction.

Fructose-1,6-bisphosphate aldolase (FBA) catalyzes an essential step in glycolysis and is present in most bacteria and eukaryotes, but conspicuously missing in Archaea and *Chlamydia*[22]. Instead, these organisms possess a different type of aldolase, which would form a complementary pattern, if not for the presence of both types of aldolases in *E. coli* and *Aquifex aeolicus*. Thus, it can be predicted that the DhnA-type enzyme functions as the only FBA in *Chlamydia* and Archaea. In this case, the prediction is buttressed by the demonstration of the FBA activity of *E. coli* DhnA protein[23].

Thymidylate synthase, an essential enzyme of DNA precursor biosynthesis is unexpectedly missing in several bacterial and archaeal species (Table 1). A complementary pattern was seen in a COG that includes uncharacterized bacterial and archaeal proteins (Table 1) whose homolog from the slime mould *Dictyostelium* has been shown to complement thymidylate synthase deficiency, although no sequence similarity to thymidylate synthases was detectable[24]. Examination of the multiple alignment of these proteins shows that the pattern of conserved residues is compatible with an enzymatic activity (data not shown), and together with the complementarity of the phylogenetic patterns with the classical thymidylate synthase, leads to the prediction that this protein is a novel thymidylate synthase unrelated to the known one. So far, *Mycobacterium tuberculosis* seems to be the only organism that encodes both types of thymidylate synthase (Table 1).

Rosetta Stone proteins

Another comparative-genomic approach that recently has received considerable attention exploits a different type of protein neighborhood by systematically analyzing protein and domain fusion (and fission)[25–27]. The basic assumption is straightforward—fusion is maintained by selection only when it facilitates kinetic coupling of consecutive enzymes in pathways or other forms of functional interaction between proteins. Therefore, those proteins that are fused in some species are likely to interact, physically or at least functionally, in other organisms. In a less prosaic language, such telling proteins fusions have been called "Rosetta Stone" proteins for they might be able to give out the mystery of the function of their components.

Rosetta Stone cases are not rare. Examination of the 2100 COGs reveals 409 unique multidomain protein architectures whose components belong to different COGs. There are a considerable number of well-known examples where a functional connection between the components of a Rosetta Stone protein is beyond doubt; for example, A and B subunits of type II topoisomerases (separate genes in bacteria, single polypeptide in eukaryotes) or RNA polymerase subunits A' and A" (single gene in bacteria and eukaryotes, separate gene in archaea). The heuristic value of the Rosetta Stone approach is, of course, not in these textbook cases, but in those where the function of at least one component is not well-understood, which is indeed true of the majority of the detected fusions. Anecdotal examination suggests that the Rosetta Stone protein sets are a mixture of readily interpretable situations, those that have potentially interesting implications, but should be approached with caution, and those that do not allow any extrapolations.

To increase the robustness of the results, the Rosetta Stone approach requires additional analytical procedures and support from other types of information. First, it is necessary to filter out "promiscuous" domains that tend to combine with a variety of other domains[25] and, in the context of the Rosetta Stone analysis, would dramatically increase the number of false-positives (examples include the DNA-binding HTH domain and the CBS domain). Second, it is important to show, as reliably as possible, that the stand-alone counterparts of a Rosetta Stone protein's components are indeed orthologs; if paralogs are involved, the fraction of false predictions is most likely to increase significantly. Last but not least, when Rosetta Stone hints are interpreted, phylogenetic patterns, results of sequence analysis, and biological considerations should be taken into account.

The complexity involved in the Rosetta Stone approach is well illustrated by the finding that in 360 of the 409 unique multidomain architectures present in the COGs, the individual components had different phylogenetic patterns. Clearly, with one of the components missing, a functional link inferred from the Rosetta Stone domain architecture is no longer relevant. Thus, while domain fusions do suggest functional association, such connections tend to be mobile in evolution, with NOGD being more of a rule than an exception.

In principle, the Rosetta Stone approach allows transitive closure (i.e., identification of not just pairs but closed sets of transitively connected components). In other words, if combinations AB, BC, and CD are detected, components A, B, C, and D are predicted to be functionally connected, perhaps forming a multisubunit complex or a pathway. Transitivity has been used, for example, in the analysis of prokaryotic signal-transduction systems, resulting in the prediction of several new signaling domains whose functions remain to be characterized experimentally[28,29]. This extension of the Rosetta Stone procedure, however, requires extra caution because the likelihood of getting on a wrong track is high[29].

Gene neighborhoods

An approach that, in many respects, is analogous to the Rosetta Stone methodology, and may be a useful complement to it, includes analysis of gene neighborhoods in genomes[26,30,31]. The central assumption,

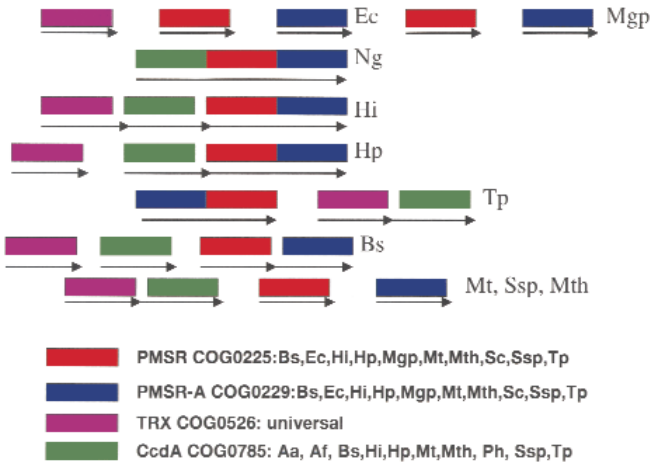

Figure 2. Phylogenetic patterns, domain fusion, and gene clustering help predict functional pathways. Color-coded rectangles indicate proteins and arrows indicate genes. Merged rectangles show fused (Rosetta Stone) proteins and merged arrows show adjacent genes (whenever genes are not adjacent, there is a gap between the arrows). For each of the proteins (domains), the COG number and the phylogenetic pattern are shown. Ng, *Neisseria gonorhoeae*; other abbreviations are as in Table 1.

again, is quite basic—functionally related genes in prokaryotes tend to form operons. The composition of operons is evolutionarily variable, so one cannot count on a particular set of functionally related genes always to comprise an operon[33]. Nevertheless, if such an operon is present in one, or better yet, in several genomes, a functional association can be predicted for other organisms, even if the corresponding genes are scattered. Furthermore, the variability of operon structure could even work to the advantage of this approach as additional functionally related genes might be occasionally drawn into operons, thus enriching the predicted network of interactions.

The catch is that prediction of unknown operons is a difficult and error-prone procedure that has never been defined in algorithmic terms. What *is* easy to do, however, is to detect adjacent or close genes provided that orthologous relationships have been identified correctly. Many of these adjacencies are, of course, functionally irrelevant, but further, detailed analysis may help predicting new functional connections.

Overbeek et al.[32] have developed an automatic procedure that detects pairs of "close" orthologs—"close" in this case is defined as belonging to the same "run" of genes (i.e., a set of genes separated by less than 300 base pairs in the respective genomes) and orthologs are operationally defined as bidirectional intergenomic best hits—and scores them according to the phylogenetic distance between the respective species, as inferred from the rRNA-based phylogenetic tree. It is expected that chance occurrence of pairs of close orthologs in phylogenetically distant genomes is much less likely than in closely related species, and accordingly, high-scoring pairs are likely to be functionally relevant. This approach allowed the successful reconstruction of several known metabolic pathways[31]. It seems that further development of this methodology, for example by considering pairs of close orthologs occurring in three or more genomes or the presence of three or more close orthologs in two genomes, could further increase the specificity and hence the predictive power of this approach.

The example in Figure 2 shows both the potential and some limitations of the Rosetta Stone approach combined with gene neighborhood information. In most organisms, protein methionyl sulfoxide reductase (PMSR) is a small, single-domain protein. However, in *Haemophilus influenzae*, *Helicobacter pylori* and *Treponema pallidum,* it is fused with another highly conserved domain (we designate it PMSR-A, after PMSR-associated) that is found as a distinct protein in all other organisms that encode PMSR (that is, the two components show the same phylogenetic pattern as immediately seen upon inspection of the corresponding COGs). Curiously, in *T. pallidum*, the order of the domains is inverted compared with *H. influenzae* and *H. pylori* (Fig. 2). The *H. influenzae* and *H. pylori* "Rosetta Stone" proteins are most closely related to each other, but the one from *T. pallidum* does not cluster with them in terms of sequence similarity, which suggests two independent fusion events. In *Bacillus subtilis*, there is no fusion, but the genes for PMSR and PMSR-A are adjacent and may form an operon (Fig. 2). Furthermore, sequence analysis of PMSR-A reveals the presence of two conserved cysteines that could be involved in oxidoreduction (E. V. Koonin, unpublished observations). All this taken together, the hypothesis that PMSR and PMSR-A are functionally and physically coupled looks extremely strong.

There is more to the story, however. In *Neisseria gonorhoeae*, a thioredoxin domain is added to the PMSR–PMSR-A fusion. Notably, in *H. influenzae*, the ortholog of this predicted thioredoxin is encoded two genes upstream of PMSR–PMSR-A. The gene in between encodes a highly conserved integral membrane protein (designated CcdA for its requirement for cytochrome *c* biogenesis in *B. subtilis*[34]), the ortholog of which is encoded next to PMSR–PMSR-A in *H. pylori* and next to thioredoxin in several other genomes (Fig. 2). Combining all this evidence from sequence analysis, phylogenetic profiles, Rosetta Stone, and gene adjacency data, it can be predicted that the PMSR, PMSR-A and thioredoxin form an enzymatic complex that catalyzes a cascade of redox reactions and is associated with the bacterial membrane via CcdA. Clearly, however, there are variations on this theme since the phylogenetic patterns for PMSR–PMSR-A and CcdA differ (Fig. 2). Furthermore, it would have been unreasonable to attempt further transitive analysis through the thioredoxin domain. First, orthologous relationships among thioredoxins are ambiguous, and second, while they are not among the most promiscuous domains, the variety of their interactions, including Rosetta Stone cases, is such that numerous false predictions would become inevitable.

Integrating DNA and protein information

The above approaches operate entirely at the level of protein sequences. A natural next step is to extract additional signals from DNA itself and combine them with information derived from proteins. This is particularly important because information contained in prokaryotic noncoding regions is insufficient to identify regulatory sites without additional data. Recently, integrated analysis of putative operators, orthologous gene sets and gene adjacency has been used to predict several regulons in bacteria and archaea[35,36].

A recent study connects the concept of phylogenetic neighborhood with experimental analysis of protein–protein interactions[36]. This is based on the premise that interactions identified using the two-hybrid system (many of which are false-positives) can be validated by showing they occur for pairs of orthologous proteins in two different species. If an interaction between proteins X and Y has been identified, for example, in yeast, and the nematode proteins X′ and Y′, which are orthologs of the respective yeast proteins, also interact, the interacting pairs X–Y and X′–Y′ are called "interologs". Conversely, if an interaction has been detected between two proteins in one species, a direct analysis of the potential interaction between their orthologs in another species is justified. Thus, a systematic, directed identification of interologs by a combination of computational and experimental means may significantly accelerate the generation of a catalog of functionally important protein–protein interactions.

The phylogenetic profile approach, the Rosetta Stone approach, and the analysis of gene clusters have been applied to complete genomes, which resulted in large sets of potentially interacting gene groups[14,25,26,32]. Eisenberg and colleagues[38] have gone further by implementing a procedure that integrates phylogenetic profiles,

Rosetta Stone results, and data on gene co-expression. Altogether, using the union of gene clusters produced by all these approaches, they detect connections for more than half of the 2,557 uncharacterized yeast proteins that have been defined as such because there was no experimental evidence as to their functions or strong sequence similarity to proteins of known functions[39].

At first glance, this may seem like an extremely impressive result, but one must realize that these connections by no means can be equated with functional prediction. First, only for 374 proteins (15% of the uncharacterized protein set), "high confidence" functional links (those based on phylogenetic profiles or on more than one prediction approach) were obtained[37]. Second, even these, purportedly most reliable predictions are plagued with ambiguity, particularly when nontrivial, specific functional inferences are involved.

A brief examination of the specific examples provided by Eisenberg and colleagues[38] gives one a glimpse of these uncertainties. The uncharacterized protein YGR021W that is highly conserved in most, if not all, completely sequenced bacterial genomes is predicted to participate in mitochondrial protein synthesis. The prediction of an essential mitochondrial function for this protein indeed seems plausible given high conservation between bacteria and eukaryotes and the presence of a likely mitochondrial import peptide in the eukaryotic members of this protein family. The validity of the connection to protein synthesis, however, can be questioned, especially because the same inference is made for the proteins of the GidA family, which are identified as "functional partners" of YGR021W. Again, a mitochondrial function for the GidA family proteins is likely, but these proteins are clearly predicted to possess an oxidoreductase activity (see, for example, COG0445), which makes the translation connection tenuous. It seems possible that the link of both these protein families to translation is, in a sense, spurious, simply reflecting the fact that many components of the mitochondrial translation machinery are highly conserved in all bacteria and eukaryotes and thus show a phylogenetic profile similar to that of YGR021W. In the same vein, among the links revealed for the translation release factor Sup35, those to translation machinery components are obvious, but the validity of the specific connections to proteins involved in protein sorting remains an open question.

Conclusions

As with most new automatic approaches in computational biology, assessing the actual power of the context-based methods for protein function prediction requires extensive testing by labor-consuming, case-by-case computational, and eventually experimental analysis. Regardless of the outcome of such assessments, it is clear that none of these methods can miraculously endow us with an "understanding" of genomes. Rather, they provide a useful extension of, and in a sense a genome-based framework for, sequence and structural analysis, which remains the cornerstone of computational genomics. This being said, the simultaneous development of these strategies by several independent groups is a sign of a new era, which is marked by explicit use of new opportunities created by the availability of a growing collection of complete *genomes*, rather than just sets of genes.

Acknowledgements

We thank L. Aravind, Arcady Mushegian, and Yuri Wolf for numerous helpful discussions of the issues considered in this article, and Martijn Huynen and Peer Bork for sending us preprints of their publications.

1. Koonin, E.V., Tatusov, R.L. & Galperin, M.Y. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**, 355-363 (1998).
2. Bork, P. et al. Predicting function: From genes to genomes and back. *J. Mol. Biol.* **283**, 707-725 (1998).
3. Bork, P. & Koonin, E.V. Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.* **18**, 313-318 (1998).
4. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html.
5. http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html.
6. http://www-fp.mcs.anl.gov/~gaasterland/genomes.html.
7. Huynen, M.J. & Snel, B. Gene and context: integrative approaches to genome analysis. *Adv. Prot. Chem.* **54**, 345–380 (2000).
8. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-106 (1970).
9. Henikoff, S. et al. Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614 (1997).
10. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
11. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33-36 (2000).
12. http://www.ncbi.nlm.nih.gov/COG. .
13. Gaasterland, T. & Ragan, M.A. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics* **3**, 199-217 (1998).
14. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U S A* **96**, 4285-4288 (1999).
15. Brown, J.R. & Doolittle, W.F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**, 456-502 (1997).
16. Makarova, K.S. et al. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**, 608-628 (1999).
17. Dandekar, T., Schuster, S., Snel, B., Huynen, M. & Bork, P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**, 115-124 (1999).
18. Huynen, M.A., Dandekar, T. & Bork, P. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol.* **7**, 281-291 (1999).
19. Ibba, M. et al. A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **278**, 1119-1122 (1997).
20. Ibba, M., Bono, J.L., Rosa, P.A. & Soll, D. Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. USA* **94**, 14383-14388 (1997).
21. Wolf, Y.I., Aravind, L., Grishin, N.V. & Koonin, E.V. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689-710 (1999).
22. Galperin, M.Y., Aravind, L. & Koonin, E.V. Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea. *FEMS Microbiol. Lett.* **183,** 269-264 (2000).
23. Thomson, G.J., Howlett, G.J., Ashcroft, A.E. & Berry, A. The dhnA gene of Escherichia coli encodes a class I fructose bisphosphate aldolase. *Biochem. J.* **331**, 437-445 (1998).
24. Dynes, J.L. & Firtel, R.A. Molecular complementation of a genetic marker in *Dictyostelium* using a genomic DNA library. *Proc. Natl. Acad. Sci. USA*. **86**, 7966–7970 (1989).
25. Marcotte, E.M. et al. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753 (1999).
26. Enright, A.J., Ilipoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).
27. Snel, B., Bork, P. & Huynen, M. Genome evolution: Gene fusion versus gene fission. *Trends Genet.* **16,** 9-11 (2000).
28. Aravind, L. & Ponting, C.P. The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* **22**, 458-459 (1997).
29. Galperin, M.Y., Natale, D.A., Aravind, L. & Koonin, E.V. A specialized version of the HD hydrolase domain implicated in signal transduction. *J. Mol. Microbiol. Biotechnol.* **1**, 303-305 (1999).
30. Doolittle, R.F. Do you dig my groove? *Nat. Genet.* **23**, 6-8 (1999).
31. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* http://www.bioinfo.de/isb/1998/01/0009/(1998).
32. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896-2901 (1999).
33. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324-328 (1998).
34. Schiott, T., Throne-Holst, M. & Hederstedt, L. Bacillus subtilis CcdA-defective mutants are blocked in a late step of cytochrome c biogenesis. *J. Bacteriol.* **179**, 4523-4529 (1997).
35. Mironov, A.A., Koonin, E.V., Roytberg, M.A. & Gelfand, M.S. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* **27**, 2981-2989 (1999).
36. Gelfand, M.S., Koonin, E.V. & Mironov, A.A. Prediction of transcription regulatory sites in Archaea by a comparative-genomic approach. *Nucleic Acids Res.* **28**, 695-705 (2000).
37. Walhout, A.J. et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287**, 116-122 (2000).
38. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).
39. Mewes, H.W., Hani, J., Pfeiffer, F. & Frishman, D. MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.* **26**, 33-37 (1998).
40. Doolittle, W.F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307-311 (1998).
41. Doolittle, W.F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).
42. Mushegian, A.R. & Koonin, E.V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93**, 10268-10273 (1996).
43. Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R. & Koonin, E.V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442-444 (1998).